Exploring self-supervised pre-trained model for surgical video understanding

Supervisors: Dr. Yasmina Al Khalil (y.al.khalil@tue.nl) and Yiping Li, (y.li9@tue.nl)

Introduction

Medical Foundation models have shown success in various applications, including disease diagnosis and text report generation. There is emerging research in this field (see Figure 1), with models being released via Large-scale Self-supervised Pre-train. The purpose of large-scale self-supervised training is to be adapted to downstream tasks via fine-tuning on smaller datasets by serving as the backbone.

The NWO-funded INTRA-SURGE project aims to develop machine-learning models with surgical video data to provide surgeons with assistance during Robot-assisted minimally invasive esophagectomy (RAMIE) surgeries (see Figure 2). As part of the INTRA-SURGE project, we aim to investigate whether endoscopic foundation models can improve robotic surgical video understanding, specifically for the surgical phase recognition task. Surgical phases refer to the steps that need to be completed in a surgery, and identifying different phases is crucial for improving surgical assistance, allowing real-time monitoring to help surgeons anticipate and adapt to critical transitions.

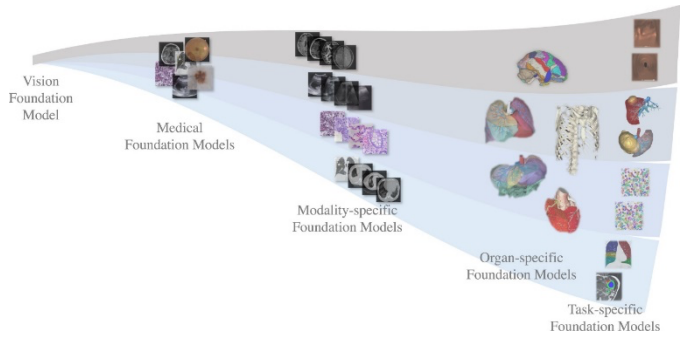

Figure1. Image from "S. Zhang and D. Metaxas. On the Challenges and Perspectives of Foundation Models for Medical Image Analysis. Medical Image Analysis"

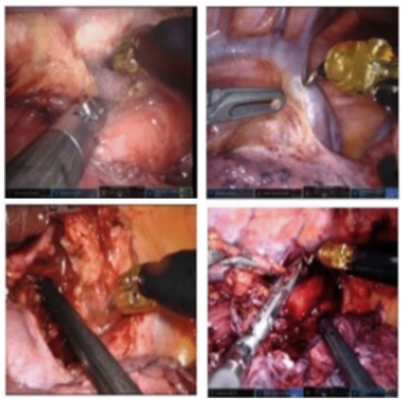

Figure2. Example Robot-assisted minimally invasive esophagectomy (RAMIE) frames

Project Goal

The project aims to use a publicly available pre-trained model[1] in the endoscopic domain to perform downstream tasks by fine-tuning. This model has been developed and proven to perform well in downstream tasks such as classification, segmentation and detection.

In this project, we want to use this pre-trained model to test its performance on the surgical phase recognition task. The students will start experimenting with the public dataset Cholec80[3] and may switch to private datasets provided by the group at a

later stage. We would like to fine-tune the model well in terms of learning rate, batch size, epoch number, feature dimension, number of temporal layers, etc.

This project is most suited for students with an interest in deep learning and programming - it will offer students a better understanding of pre-training and fine-tuning in a practical way, providing valuable hands-on experience with deep learning, as well as opportunities to refine programming, writing and presentation skills.

Contact

Yiping Li, y.li9@tue.nl

Dr. Yasmina Al Khalil, y.al.khalil@tue.nl

References

[1] Wang, Z., Liu, C., Zhang, S., & Dou, Q. (2023, October). Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 101-111). Cham: Springer Nature Switzerland.

[2] Ramesh, S., Srivastav, V., Alapatt, D., Yu, T., Murali, A., Sestini, L., ... & Padoy, N. (2023). Dissecting self-supervised learning methods for surgical computer vision. Medical Image Analysis, 88, 102844.

[3] A.P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, N. Padoy, EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos, IEEE Transactions on Medical Imaging (TMI), arXiv preprint, 2017

[4] OpenMEDLab (github.com)